Congresso Brasileiro de Engenharia Clínica – CBECLIN 2026
21 de maio de 2026

Avaliação Crítica de Métricas em *Deep Learning* para Classificação de Tumores Cerebrais por RM: Limitações da Acurácia e Implicações para SaMD

MARTINS, C. S. G.[1]; DANTONA, A. F. S. [2*]
[1] Ministério da Saúde, SCTIE/CGDIM, Brasília/Brasil
[2] TECSAÚDE, Recife/Brasil

Desafio e Objetivo

A aplicação de inteligência artificial na medicina tem avançado rapidamente, com modelos de *deep learning* frequentemente alcançando acurácias superiores a 95% na classificação de tumores cerebrais por ressonância magnética (RM). Esse desempenho, embora expressivo, tem sido utilizado tanto em apresentações de fornecedores como critério positivo para a incorporação, quanto na documentação de softwares como dispositivo médico (SaMD), conforme classificação da ANVISA pela RDC 657/2022 [1]. Entretanto, essa interpretação simplificada desconsidera limitações críticas da acurácia como métrica isolada.

Ademais, a acurácia global pode mascarar desempenhos desiguais entre classes, ocultar taxas de falso negativo clinicamente inaceitáveis e oferecer falsa segurança quanto à generalização. Mesmo em datasets com desbalanceamento discreto, a métrica tende a refletir predominantemente o desempenho das classes majoritárias, subestimando erros em classes minoritárias [2], que frequentemente correspondem aos diagnósticos de maior relevância clínica. Diretrizes internacionais, como a da The Lancet Digital Health [3] e da European Society of Medical Imaging Informatics [4], destacam que métricas isoladas não representam adequadamente o desempenho em ambiente real e podem superestimar a utilidade clínica de sistemas de IA.

Nos serviços de saúde, o engenheiro clínico é frequentemente o profissional que emite parecer técnico sobre aquisição de tecnologias. Com a crescente incorporação de SaMD baseados em IA, esse papel exige capacidade de avaliar criticamente as métricas de desempenho, identificando quais resultados e condições de validação são representativos do contexto clínico de uso pretendido.

Hipótese: a acurácia global elevada, tomada isoladamente como evidência suficiente de desempenho, pode ser inadequada na avaliação de modelos de IA em saúde.

Objetivo: avaliar criticamente a suficiência da acurácia global como métrica isolada na avaliação de dois modelos de deep learning aplicados à classificação de tumores cerebrais por RM, à luz do desempenho por classe, do risco dos erros de classificação e da representatividade do dataset.

Resumo da Solução Adotada

Modelos de *deep learning* são sistemas computacionais que aprendem a reconhecer padrões em imagens a partir de um grande número de exemplos rotulados, sem que regras de reconhecimento sejam programadas manualmente. No contexto de imagem médica, esses modelos são treinados para associar características visuais, como formato, textura e intensidade de sinal, a categorias diagnósticas previamente definidas [5].

Dois paradigmas principais são utilizados neste tipo de tarefa. No primeiro, uma rede é construída e treinada do zero, aprendendo exclusivamente a partir das imagens do problema em questão. No segundo, denominado *transfer learning* (aprendizado por transferência), parte-se de um modelo previamente treinado em um vasto acervo de imagens naturais (não médicas), tipicamente o ImageNet, com mais de um milhão de fotografias e mil categorias, cujas camadas já codificam detectores genéricos de bordas, texturas e formas. Esse modelo é então adaptado para a tarefa médica específica. A vantagem prática é expressiva: o modelo de partida já "sabe ver" estruturas visuais complexas, exigindo menos dados e menos tempo de treinamento para alcançar bom desempenho. A comparação entre os dois paradigmas permite avaliar quanto o conhecimento prévio codificado no modelo base contribui para o desempenho final e quais as implicações disso para a avaliação de SaMD.

O estudo implementou e comparou dois modelos de *deep learning* para classificação automática de tumores cerebrais em imagens de RM, utilizando o Brain Tumor MRI Dataset [6], conjunto público disponível na plataforma Kaggle, compilado por Nickparvar a partir de dados originalmente publicados por Cheng et al. [7] e Chakrabarty [8]. O *dataset* contém quatro classes: glioma (1.321 imagens de treino, 300 de teste), meningioma (1.339/306), sem tumor (1.595/405) e tumor pituitário (1.457/300), totalizando 5.712 imagens de treino e 1.311 de teste. A razão de desbalanceamento entre a classe mais representada (sem tumor) e a menos representada (glioma) é de 1,21x. O *dataset* não disponibiliza informações sobre idade, sexo ou etnia dos pacientes, nem sobre fabricante do equipamento de RM, intensidade de campo magnético, protocolo de aquisição ou sequência utilizada.

Modelo 1: CNN baseline. Rede convolucional construída do zero com três blocos convolucionais (32, 64 e 128 canais), BatchNorm, ReLU, MaxPooling, Global Average Pooling e classificador com Dropout 0,5. Total de 288.420 parâmetros treináveis. Treinada por 20 épocas com Adam (lr=0,001), CrossEntropyLoss com pesos de classe balanceados e ReduceLROnPlateau.

Modelo 2: EfficientNet-B3 com transfer learning. Arquitetura pré-treinada no ImageNet (10.702.380 parâmetros totais), com treinamento em duas fases: (a) feature extraction (10 épocas, backbone congelado, apenas 6.148 parâmetros treináveis no classificador) e (b) fine-tuning (15 épocas, *backbone* descongelado com taxa de aprendizado diferenciado: 1e-5 para backbone, 1e-4 para classificador) [9].

Pré-processamento e *data augmentation*. Imagens redimensionadas para 224x224 (*baseline*) e 300x300 (*transfer learning*), com normalização pelos parâmetros do ImageNet.

Data augmentation no treino: inversão horizontal, rotação aleatória (±15°) e ColorJitter (brilho e contraste ±0,2). Validação e teste sem *augmentation*.

Tratamento do desbalanceamento. Pesos de classe calculados com compute_class_weight (sklearn), resultando em pesos entre 0,90 (sem tumor) e 1,09 (glioma), aplicados à função de perda CrossEntropyLoss.

Resultados. A CNN baseline alcançou 92,79% de acurácia de validação. O EfficientNet-B3 atingiu 96,94% na validação e 97,48% no teste, com desempenho por classe (proporção de acertos, equivalente ao *recall*) variando entre 94,4% (meningioma) e 99,8% (sem tumor). A matriz de confusão revelou 17 confusões entre glioma e meningioma no conjunto de teste, correspondendo à maior fonte de erro do modelo.

Posicionamento analítico. Diferentemente de abordagens que se limitam a comparar acurácias, este estudo adota perspectiva crítica sobre a adequação das métricas utilizadas, as limitações do dataset e os riscos clínicos associados à interpretação simplista de resultados aparentemente elevados. A contribuição central não reside na comparação entre modelos, mas na análise das condições sob as quais métricas de desempenho podem ser inadequadamente interpretadas em contextos de saúde, e no que isso implica para o engenheiro clínico que avalia, adquire e gerencia SaMD em estabelecimentos de saúde.

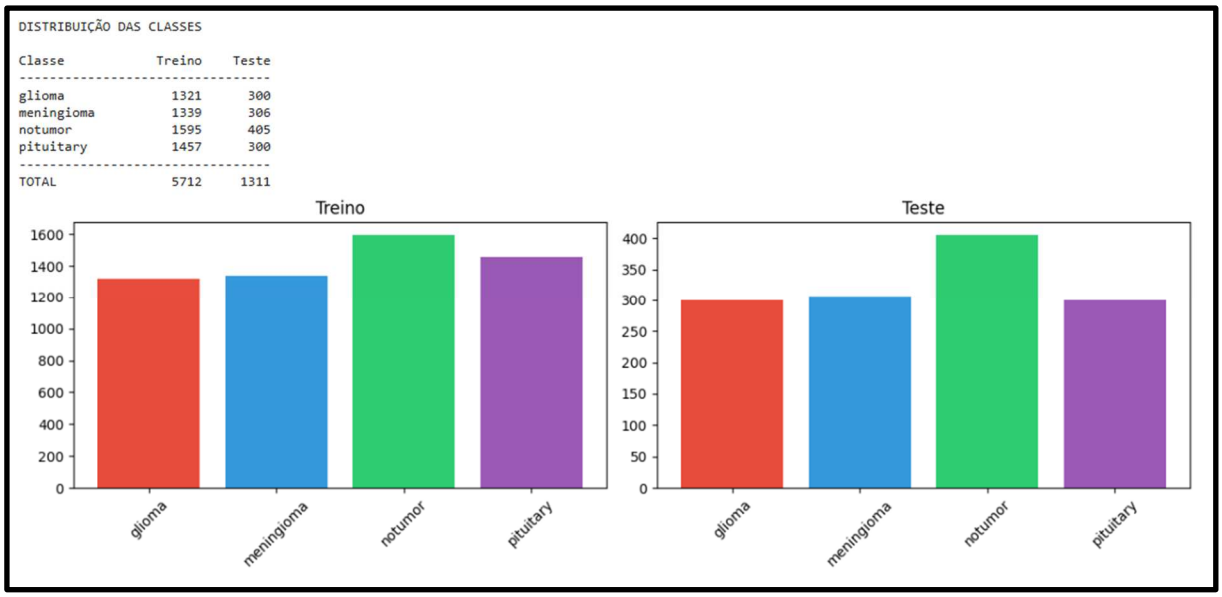


Figura 1: Amostras de imagens de RM por classe

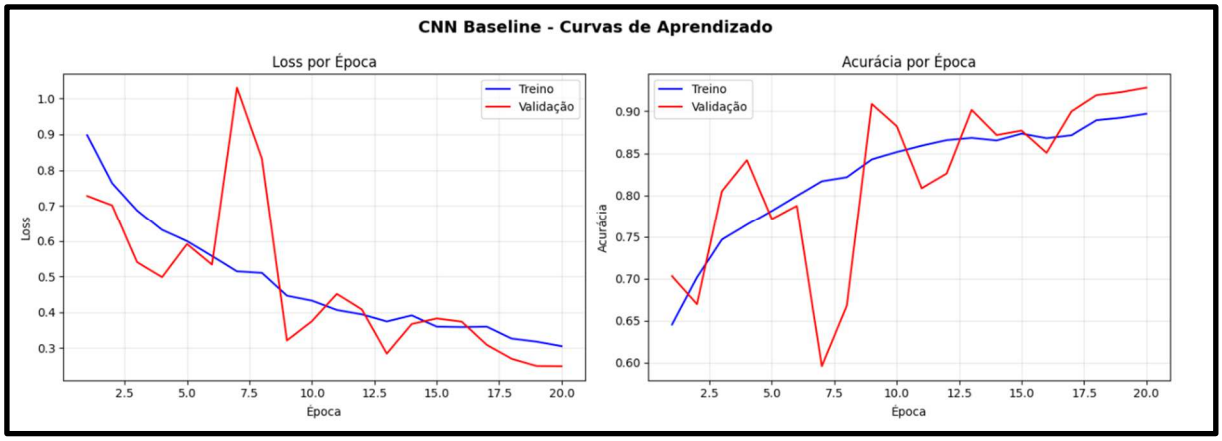


Figura 2: Curvas de aprendizado da CNN baseline e do EfficientNet-B3 (*loss* e acurácia por época)

Método Utilizado

Estudo experimental com abordagem analítico-crítica, baseado na implementação, treinamento e avaliação de dois modelos de deep learning para classificação de tumores cerebrais por RM.

Base de Dados (*Dataset*): Brain Tumor MRI Dataset [6], a partir de dados de Cheng et al. [7] e Chakrabarty [8], disponível publicamente na plataforma Kaggle. Contém 7.023 imagens de RM cerebral em quatro classes (glioma, meningioma, sem tumor, tumor pituitário). Divisão original: 5.712 treino (subdividido em 90% treino efetivo e 10% validação, 80%/20% na fase *transfer learning*) e 1.311 testes.

Razão de desbalanceamento: 1,21x.

Limitações: O *dataset* não fornece metadados clínicos (idade, sexo, etnia) nem técnicos (modelos de RM, intensidade de campo, tipo de gradiente, protocolo de aquisição).

Modelos: (a) CNN baseline com três blocos convolucionais (288.420 parâmetros, 20 épocas); (b) EfficientNet-B3 pré-treinada no ImageNet (10,7 milhões de parâmetros, 25 épocas em duas fases). Ambos utilizaram CrossEntropyLoss com pesos de classe, otimizador Adam e ajuste da taxa de aprendizado com ReduceLROnPlateau.

Métricas reportadas: acurácia global (validação e teste), proporção de acertos por classe (equivalente ao recall) e matriz de confusão.

Abordagem analítica: os resultados foram analisados criticamente sob cinco dimensões: (i) impacto do desbalanceamento de dados, (ii) limitações da acurácia como métrica isolada, (iii) riscos clínicos dos erros de classificação, (iv) ausência de caracterização populacional e técnica do dataset, e (v) implicações para avaliação de SaMD pelo engenheiro clínico. Essa análise foi conduzida à luz dos requisitos regulatórios da ANVISA (RDC 657/2022 [1], RDC 509/2021 [10]), da Resolução CFM 2.454/2026 [11] e da literatura sobre validação de modelos preditivos em saúde [12].

Principais Resultados Encontrados

A Tabela 1 apresenta a distribuição do dataset. A Tabela 2 apresenta o desempenho dos modelos.

Tabela 1. Distribuição do dataset por classe (Brain Tumor MRI Dataset [6]).

Classe	Treino	Teste	% do total
Glioma	1.321	300	23,1
Meningioma	1.339	306	23,4
Sem tumor	1.595	405	28,5
Pituitário	1.457	300	25,0
Total	**5.712**	**1.311**	**100**

Tabela 2. Desempenho dos modelos.

Modelo	Parâmetros	Validação	Teste	Menor desempenho (teste)
CNN baseline	288 mil	92,79%	—	—
EfficientNet-B3	10,7 M	96,94%	97,48%	Meningioma: 94,4%

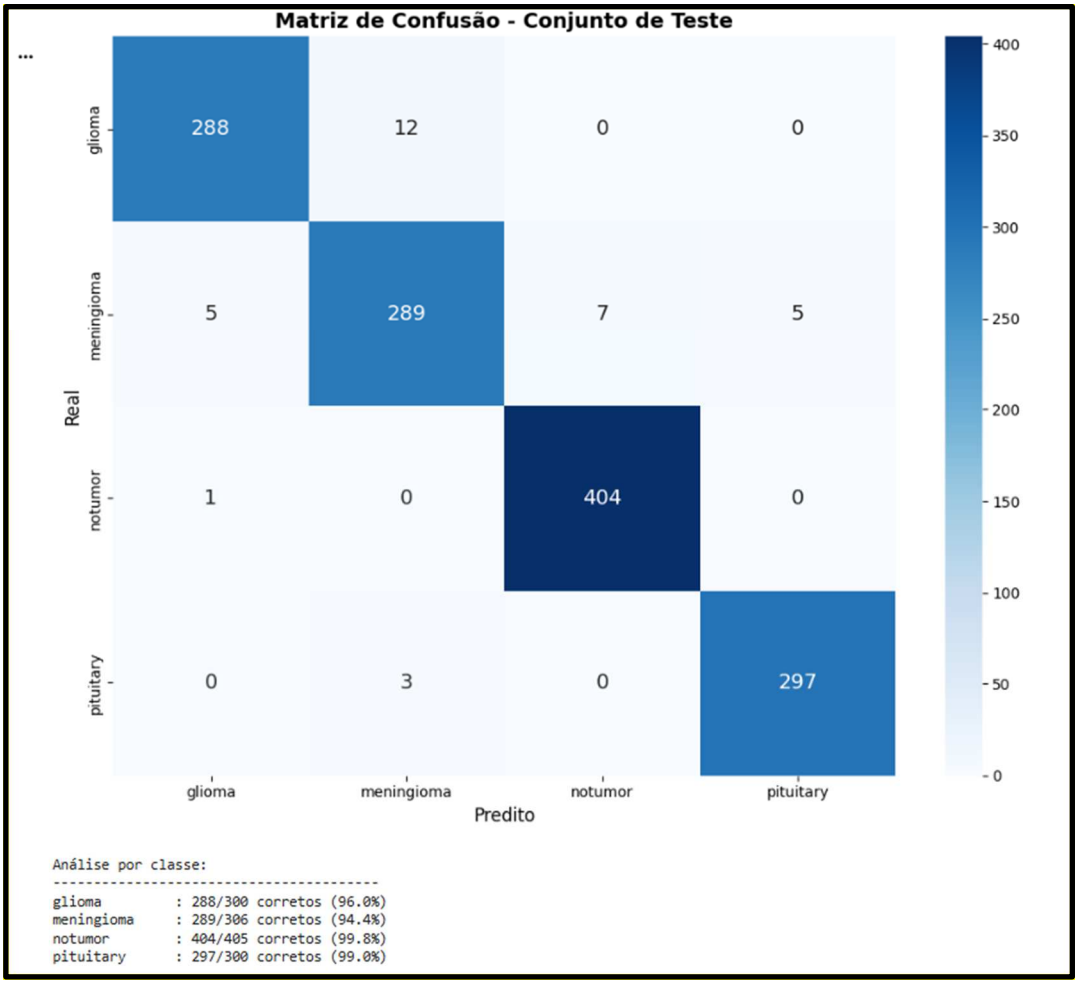
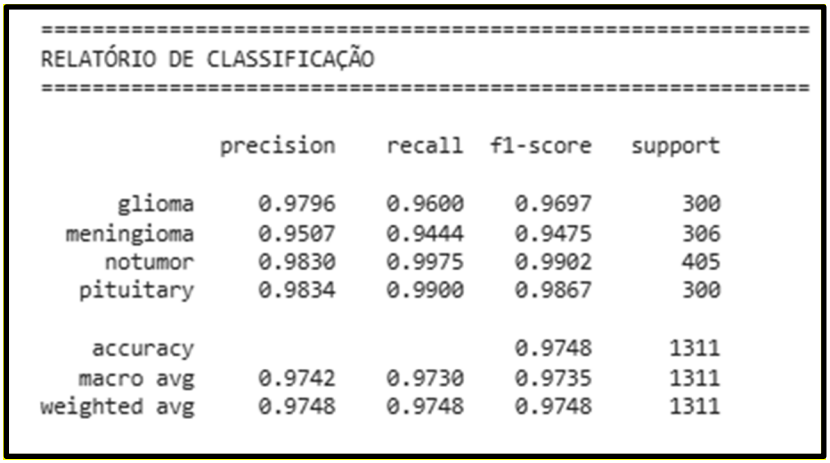


Figura 3: Matriz de confusão do EfficientNet-B3 no conjunto de teste

A matriz de confusão do EfficientNet-B3 revelou 12 gliomas classificados como meningioma e 5 meningiomas como glioma, totalizando 17 confusões entre essas duas classes, que respondem pela maior concentração de erros do modelo.

A seguir, a Figura 4 demonstra que a acurácia global, quando analisada isoladamente, não traduz integralmente o comportamento do modelo por classe, uma vez que, apesar do valor global elevado, o desempenho não é homogêneo entre as classes, com menor *recall* e *F1-score* para meningioma.



```
============================================================
RELATÓRIO DE CLASSIFICAÇÃO
============================================================

              precision    recall  f1-score   support

      glioma     0.9796    0.9600    0.9697       300
  meningioma     0.9507    0.9444    0.9475       306
     notumor     0.9830    0.9975    0.9902       405
   pituitary     0.9834    0.9900    0.9867       300

    accuracy                         0.9748      1311
   macro avg     0.9742    0.9730    0.9735      1311
weighted avg     0.9748    0.9748    0.9748      1311
```

Figura 4. Métricas de desempenho por classe do EfficientNet-B3 no conjunto de teste.

Desbalanceamento de dados. Embora o desbalanceamento do dataset seja leve (1,21x), a classe com maior representação (sem tumor, 28,5%) alcançou a maior proporção de acertos (99,8%), enquanto a classe com menor representação relativa (glioma, 23,1%) apresentou desempenho inferior (96,0%). O uso de pesos de classe atenua, mas não elimina o viés.

Limitações da acurácia. A acurácia isoladamente não é suficiente para avaliação de modelos em saúde, especialmente em cenários com múltiplas classes ou desbalanceamento, nos quais métricas como precisão, recall e F1-score fornecem uma visão mais completa do desempenho [13]. As recomendações da EuSoMII [4] estabelecem que sensibilidade, especificidade e AUROC devem ser complementadas por métricas dependentes de prevalência, como precisão, F1-score e Matthews Correlation Coefficient (MCC), para garantir uso clínico seguro. Marra [14] demonstrou que o AUROC, frequentemente adotado como alternativa à acurácia, também apresenta limitações importantes em datasets desbalanceados, propondo a família de métricas balanceadas (G4, P4 e MCC) como alternativas mais robustas para validação de classificadores binários em dispositivos médicos. A Tabela 3 sintetiza as principais limitações e indica o que o engenheiro clínico deve exigir do fabricante ao avaliar um SaMD.

Tabela 3. Limitações da acurácia e informações a exigir do fabricante de SaMD.

Limitação da acurácia	Implicação clínica	O que exigir do fabricante
Não discrimina erros por classe	Oculta desempenho inferior em diagnósticos críticos	Sensibilidade e especificidade por classe
Não diferencia FP de FN	Impede avaliar gravidade do tipo de erro	Matriz de confusão completa e análise de erros
Inflada por classes majoritárias	Superestima desempenho global	F1-score por classe e F1 macro
Não captura confiança da predição	Predições incertas tratadas como certas	Calibração do modelo e curva ROC por classe
Insensível a variações populacionais	Desempenho pode degradar em outras populações	Caracterização do dataset (idade, sexo, equipamento, protocolo)
Não avalia estabilidade	Desempenho pode variar entre centros	Validação externa em dataset independente

No presente estudo, a acurácia global de 97,48% coexiste com 94,4% para meningioma, diferença de 3 pontos percentuais que, em escala populacional, representa número significativo de classificações incorretas. A análise detalhada de sensibilidade, especificidade e F1-score por classe deve acompanhar a avaliação completa do modelo e de seu contexto de aplicação.

Risco clínico. A Tabela 4 apresenta o impacto clínico dos erros de classificação.

Tabela 4. Impacto clínico dos erros de classificação em tumores cerebrais.

Tipo de erro	Exemplo neste estudo	Consequência clínica
Falso negativo	1 meningioma classificado como sem tumor	Atraso diagnóstico, progressão tumoral, perda de janela terapêutica
Confusão entre tipos	12 gliomas classificados como meningioma	Conduta terapêutica inadequada, planejamento cirúrgico incorreto, protocolo oncológico equivocado
Falso positivo	1 sem tumor classificado como glioma	Exames invasivos desnecessários, ansiedade do paciente, custos ao sistema de saúde

Os 17 erros de confusão entre glioma e meningioma são particularmente relevantes porque esses tumores possuem prognósticos, abordagens cirúrgicas e condutas distintas. Uma classificação errônea pode direcionar o paciente para protocolo inadequado.

Limitações do dataset. A ausência de caracterização populacional limita a generalização do modelo. O Brain Tumor MRI Dataset [6] não informa idade, sexo ou etnia dos pacientes, protocolos e especificações do equipamento de RM. Essas variáveis influenciam diretamente a aparência das imagens e, consequentemente, o desempenho do modelo. Um classificador treinado predominantemente em imagens de equipamentos 1,5T pode apresentar degradação ao processar imagens de 3T, e vice-versa. Sem essa informação, a reprodutibilidade

e a generalização do modelo para outros centros e populações permanecem indeterminadas. Ruitenbeek et al. [15] demonstraram empiricamente esse risco ao validar, em coorte holandesa, uma ferramenta de IA treinada em dados indianos para detecção de fraturas em radiografias: embora o desempenho de classificação tenha se mostrado robusto (AUC de 0,92), a localização das fraturas variou de 7% a 90% dependendo da região anatômica, evidenciando que bom desempenho em um contexto de treinamento não garante generalização para outras populações ou anatomias na aplicação de IA para um mesmo tipo de exame.

SaMD e o papel do engenheiro clínico. Nos processos de avaliação e incorporação de tecnologias em serviços de saúde, cabe ao engenheiro clínico examinar criticamente se as evidências apresentadas pelo fabricante são de fato representativas do uso pretendido. Em consonância com a Resolução CFM 2.454/2026 [11], essa análise não deve se limitar a métricas agregadas de desempenho, devendo contemplar, no mínimo, transparência quanto às condições de validação, limitações conhecidas, desempenho por classe e, quando pertinente, por subgrupo, bem como matriz de confusão, caracterização do dataset e validação externa, sem prejuízo de outros requisitos relevantes à avaliação da tecnologia, como análise de riscos e aderência ao contexto de aplicação, que não constituem o foco específico deste estudo.

A avaliação de modelos de IA em dispositivos médicos exige critérios complementares aos princípios fundamentais de validação de dispositivos médicos. A fragilidade na validação de SaMD transcende especialidades, como demonstrado por revisões sistemáticas em outros domínios da imagem médica [16][17].

Conclusão. Os modelos avaliados alcançaram acurácias de 92,79% e 97,48%, resultados que isoladamente podem sugerir capacidade de classificação satisfatória do modelo. Contudo, a análise demonstrou que a acurácia global não representa integralmente o comportamento do algoritmo classificador, ao mascarar diferenças relevantes entre classes e concentrar erros entre glioma e meningioma, tipos tumorais com condutas terapêuticas distintas. A ausência de caracterização populacional do *dataset* também limita a avaliação de vieses e da capacidade de generalização do modelo. Para o profissional, a principal contribuição deste estudo é que a avaliação de SaMD baseados em IA não pode se restringir à acurácia. Em processos de avaliação e incorporação, a interpretação da acurácia global requer exame conjunto com outras métricas de desempenho e com as condições de validação, conforme o contexto de aplicação. Nesse cenário, conhecimentos de métricas para validação de modelos preditivos passam a integrar o repertório técnico necessário à avaliação de tecnologias baseadas em IA em serviços de saúde.

Referências

[1] ANVISA. Resolução da Diretoria Colegiada RDC 657, de 24 de março de 2022. Dispõe sobre a regularização de software como dispositivo médico (SaMD). Diário Oficial da União, Brasília, 2022.

[2] HE, H.; GARCIA, E. A. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, v. 21, n. 9, p. 1263-1284, 2009. DOI: 10.1109/TKDE.2008.239.

[3] McCAGUE, C. et al. Position statement on clinical evaluation of imaging AI. The Lancet Digital Health, v. 5, p. e400–e402, jul. 2023. DOI: 10.1016/S2589-7500(23)00090-0.

[4] KLONTZAS, M. E. et al. ESR Essentials: common performance metrics in AI — practice recommendations by the European Society of Medical Imaging Informatics. European Radiology, v. 36, p. 1528–1540, 2026. DOI: 10.1007/s00330-025-11890-w.

[5] LeCUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, v. 521, n. 7553, p. 436–444, 2015. DOI: 10.1038/nature14539.

[6] NICKPARVAR, M. Brain Tumor MRI Dataset. Kaggle, 2021. Disponível em: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. Acesso em: nov. 2025.

[7] CHENG, J. et al. Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Fine-Tuning. PLoS ONE, v. 10, n. 10, e0140381, 2015. DOI: 10.1371/journal.pone.0140381.

[8] CHAKRABARTY, N. Brain MRI Images for Brain Tumor Detection. Kaggle, 2019. Disponível em: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection. Acesso em: nov. 2025.

[9] TAN, M.; LE, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019. p. 6105-6114.

[10] ANVISA. Resolução da Diretoria Colegiada RDC 509, de 27 de maio de 2021. Dispõe sobre os requisitos para a gestão de tecnologias em saúde em estabelecimentos de saúde. Diário Oficial da União, Brasília, 2021.

[11] BRASIL. Conselho Federal de Medicina. Resolução CFM nº 2.454/2026. Dispõe sobre a utilização de inteligência artificial na prática médica. Diário Oficial da União, Brasília, 2026.

[12] PARK, S. H.; HAN, K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology, v. 286, n. 3, p. 800-809, 2018. DOI: 10.1148/radiol.2017171920.

[13] SOKOLOVA, M.; LAPALME, G. A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing and Management, v. 45, n. 4, p. 427-437, 2009. DOI: 10.1016/j.ipm.2009.03.002.

[14] MARRA, A. G4 & the balanced metric family — a novel approach to solving binary classification problems in medical device validation & verification studies. BioData Mining, v. 17, n. 43, 2024. DOI: 10.1186/s13040-024-00402-z.

[15] RUITENBEEK, H. C. et al. Cross-validation of an artificial intelligence tool for fracture classification and localization on conventional radiography in Dutch population. Insights into Imaging, v. 16, n. 150, 2025. DOI: 10.1186/s13244-025-02034-1.

[16] DASHTI, M. et al. Use of artificial intelligence for detection of MB2 canals in maxillary first molars on CBCT: a systematic review and meta-analysis. BMC Oral Health, v. 25, n. 1860, 2025. DOI: 10.1186/s12903-025-07254-x.

[17] GHADERZADEH, M.; GARAVAND, A.; SALEHNASAB, C. Artificial intelligence in polycystic ovary syndrome: a systematic review of diagnostic and predictive applications. BMC Medical Informatics and Decision Making, v. 25, n. 427, 2025. DOI: 10.1186/s12911-025-03255-6.

Agradecimentos

Ao Ministério da Saúde (CGDIM/DECIS/SCTIE) pelo suporte institucional. Aos autores do Brain Tumor MRI Dataset [6] e dos datasets originários [7,8] pela disponibilização pública dos dados utilizados neste estudo.